Transportation and Traffic Theory 2009
Edited by W.H.K. Lam, S.C. Wong and H.K. Lo

Cordon Pricing Consistent with the Physics of Overcrowding

Nikolas Geroliminis and David M. Levinson, University of Minnesota, U.S.A.

Abstract This paper describes the modeling of recurring congestion in a network. It is shown that the standard economic models of marginal cost cannot describe precisely traffic congestion in networks during time-dependent conditions. Following a macroscopic traffic approach, we describe the equilibrium solution for a congested network in the no-toll case. A dynamic model of cordon-based congestion pricing (such as for the morning commute) for networks is developed consistent with the physics of traffic. The paper combines Vickrey's theory with a macroscopic traffic model, which is readily observable with existing monitoring technologies. The paper also examines some policy implications of the cordon-based pricing to treat equity and reliability issues, i.e. in what mobility level a city should choose to operate. An application of the model in a downtown area shows that these schemes can improve mobility and relieve congestion in cities.

1. Introduction

Traffic congestion is a classic externality, increasingly pervasive in urban areas. Constructing new infrastructure is an expensive solution to decrease congestion, not only because of the tremendous cost to keep pace with population increases and the resulting increase in travel demand but also because of the phenomenon of induced demand. To alleviate traffic congestion in cities, road pricing has been proposed by many researchers as an effective policy. The intention is to alter travelers' behavior enough to reduce congestion by charging them for the externalities they create. This paper aims to lay out a clear model of traffic congestion that can be applied to cordon pricing, and allows us to critique the plausibility of several economic models of congestion that have appeared in the literature.

The vast literature of congestion pricing can be composed in two categories, marginal-cost pricing models and bottleneck models. The theoretical background of marginal-cost pricing has relied on the fundamental concept, first introduced by Pigou (1920) and followed by Vickrey (1963), Beckmann (1965) and other re-

searchers; If on each link of a network a toll is charged, which equals the additional congestion cost imposed on other users by an extra traveler, the sum of consumers' surplus and total revenue is maximized. In the traffic assignment literature tolls of this type (first-best pricing) have been proposed to drive a user equilibrium pattern (Wardrop 1952) toward a system optimum. Despite their idealized theoretical basis, first-best pricing models have been impractical and difficult to implement. Marchand (1968) studied the theory of second-best tolls using a general equilibrium model for network with two routes in parallel, for a fixed period. According to the second-best pricing models, e.g. Arnott et al. (1990a) and McDonald (1995), tolls are charged in a subset of selected links, where are the bottlenecks.

Instead of charging in individual separate links, cordon- or area-based pricing schemes has been developed and applied in different cities including Singapore, Trondheim, Oslo, Stockholm, and London. Recently, Maruyama and Sumalee (2007) compared performance of cordon- and area-road pricing schemes on their efficiency and equity. Congestion pricing in general networks has been discussed mainly for the static situation. In network level, Yang and Huang (1998) examined the principle of marginal-cost pricing in a road network, Anderson and Mohring (1997) examined congestion on the Twin Cities road network having drivers face marginal rather than average costs to reflect optimal prices using a user equilibrium assignment for a single period, while Liu and Boyce (2002) examined the condition for dynamic traffic assignment with multiple time periods, and Yang and Lam (1996) analyzed the optimal road tolls under conditions of both queuing and congestion. The basic ambiguity in most of these models is that traffic conditions are considered stationary or semi-stationary. One of the aspects of our paper is to analyze, based on a new methodology, dynamic pricing schemes (such as used in Stockholm, and to a lesser extent in Singapore) that allow system managers to trade off equity vs. reliability.

For the dynamic case, bottleneck models, in contrast with the marginal-cost pricing models, treat the demand and departure time decision endogenously. To the best of our knowledge, the original contribution is Vickrey (1969). This model deals with the time-dependent equilibrium distribution of arrivals at a single bottleneck, with particular reference to the morning commute. A traveler usually experiences a delay cost of waiting in the queue and a penalty, "schedule delay", which is the difference between the actual time passing the bottleneck and the desired time; accordingly, he may adjust his departure time to avoid highly congested periods. Equilibrium obtains when no individual has an incentive to alter his departure time. Vickrey's model has been extended and analyzed in various ways. Small (1982) considered the endogenous scheduling of work trips. Smith (1984) and Daganzo (1985) proved the existence and uniqueness of Nash equilibrium if all travelers experience the same convex cost function. Newell (1987) analyzed the morning commute problem for non-identical travelers, while Arnott et al. (1993) extended the model for the case of elastic demand. Arnott et al. (1990b) used Vickrey's model to evaluate toll policies (fine toll vs. coarse toll) for a common desired arrival time for all travelers.

The study of dynamic congestion pricing for a network has been limited to some idealized situations due to the complications of dynamic traffic assignment and network structure and the tediousness of an analytical approach. Huang and Yang (1996) used control theory to determine the optimal variable tolls on a congested network of parallel routes with elastic demand, while Yang and Meng (1998) applied a space-time expanded network approach (initially proposed by Ford and Fulkerson (1962)) to estimate optimal toll in a queueing network with elastic demand. The missing piece for these analyses is that, as we will explain later, the average flow and the output rate of a network, i.e. the capacity of bottlenecks decreases when there are spillovers from downstream bottlenecks, typically found with high values of car density. Lago and Daganzo (2007) analyzed the morning commute problem for some simple freeway networks with queue spillovers and merge interactions. For a more extended review of the different types of pricing problems (e.g. first- or second-best pricing, dynamic road pricing, bottleneck models etc.) the reader can refer to Yang and Huang (2005).

The traditional network supply curve (desired or input demand vs. average travel cost) is not consistent with the physics of traffic. This is because for a given average flow (i.e. desired demand over a period of time) the total cost (expressed in delay terms) (i) is sensitive, during congested conditions, to small variations of flow within the given period and (ii) depends on the initial state of the system (the level of congestion). It has been broadly shown through simulation and field experiments (for example Geroliminis and Daganzo (2008) plots between pertinent variables (flow, speed, delay) on a spatially disaggregated level, i.e. in one link in a network, are very chaotic and do not follow a well-defined curve. (The main reason is that, at a link level, traffic systems are not in steady-state conditions.) Thus, the estimated congestion toll based on idealized versions of these curves may not be optimal and the system may be either still congested (if under-priced) or very uncongested (if over-priced).

Instead, in this paper we propose a cordon-based congestion pricing scheme which (i) is easier to implement in real cities because it should have lower collection and transaction costs than link-based or area-based tolling (Levinson 2002; Levinson and Odlyzko 2008); (ii) is based on traffic models that are readily observable with existing monitoring technologies, they can be verified and their predictions trusted; and (iii) can account for dynamic characteristics of traffic. We propose a dynamic model of congestion pricing (such as for the morning commute) of congested networks consistent with the physics of traffic. This paper combines Vickrey's theory with a macroscopic traffic model, which has been recently proposed and tested (Daganzo 2007; Geroliminis and Daganzo 2007, 2008). According to this, traffic in large urban regions (neighborhoods) can be modeled dynamically at an aggregate level, if the neighborhoods are uniformly congested.

The remainder of the paper is structured as follows: Section 2 introduces the most important features of the macroscopic fundamental diagram (MFD) and shows the ambiguity of the marginal-cost models to deal with dynamic conditions. Section 3 introduces the equilibrium conditions for a network governed by an MFD and gives some results from a real experiment. Section 4 estimates the op-

timal time-varying fine toll for this network and provides insights for the trade-off between equity and reliability, while Section 5 provides discussion and some future work.

2. A Macroscopic Fundamental Diagram (MFD) for Urban Traffic

It has been recently proposed and tested in Daganzo (2007) and Geroliminis and Daganzo (2007, 2008) that traffic in large urban regions can be modeled dynamically at an aggregate level, if the neighborhoods are uniformly congested. These papers showed, using a micro-simulation of the San Francisco Business district and a field experiment in downtown Yokohama (Japan), (i) that urban neighborhoods approximately exhibit a "Macroscopic Fundamental Diagram" (MFD) relating the number of vehicles (accumulation) in the neighborhood to the neighborhood's average speed (or flow) and (ii) there is a robust linear relation between the neighborhood's average flow and its total outflow (rate vehicles reach their destinations). The experiment used a combination of fixed detectors and floating vehicle probes as sensors. Fig. 1 shows some findings of the experiment in Yokohama (time resolution is 5min). It was observed that when the somewhat chaotic scatter-plots of speed vs. density from individual fixed detectors were aggregated for a 10km^2 region, the scatter nearly disappeared and points grouped neatly along a smoothly declining curve (compare Fig. 1a with 1b-1d). The same references also showed that (a) the MFD is a property of the network itself (infrastructure and control) and not of the demand, i.e. the MFD should have a well-defined maximum and remain invariant when the demand changes both with the time-of-day and across days (it may vary if the O-D pattern of demand changes significantly, though, e.g. due to an event or evacuation); (b) the space-mean flow, is maximum for the same value of density of vehicles or average speed, independent of the origin-destination tables; (c) the average trip length for the study region is about constant with time, i.e. the total outflow vs. density curve is a scaled up version of the curve in Fig. 1b; and (d) the MFD can be estimated accurately using existing monitoring technologies (e.g. detector data, GPS etc). In this paper, we utilize the properties of an MFD to develop a cordon-based congestion pricing scheme that overcomes some limitations of the existing models.

The Macroscopic Fundamental Diagram (MFD) resembles the classical Microscopic Fundamental Diagram (μFD). It has been observed from empirically derived μFD that the same flow can be achieved on many links at two different speeds. This has been dubbed the "backward-bending" phenomenon (Hau 1998; Crozet and Marlot 2001) or "hypercongestion". There are at least two sources for "backward-bending" speed-flow relationships. The first has to do with the point of observation. Observing the lane flow upstream of a bottleneck gives the impression of a backward bending relationship, but this disappears at the bottleneck it-

self. Under any given demand pattern, flow and speed are a unique pair. When demand is below the downstream active bottleneck's capacity, a flow on an upstream link can be achieved at high speed. When demand is above the downstream active bottleneck's capacity, the same flow on the upstream link can only be achieved at a low speed because of queueing. The second has to do with a capacity drop at the bottleneck itself under congested conditions. However, much research reports that this drop is slight to non-existent (Cassidy and Bertini 1999; Zhang and Levinson 2004).

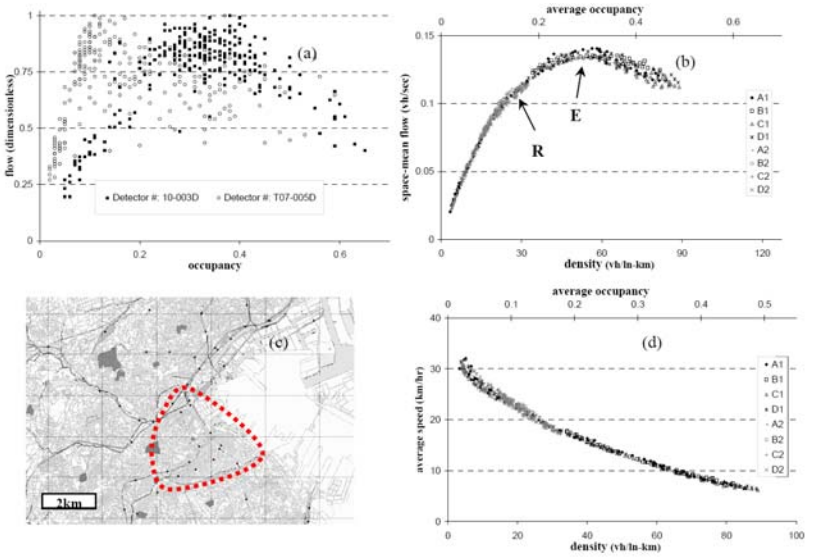

Fig. 1. Loop detector data (taken from Geroliminis and Daganzo 2008): (a) flow vs. occupancy pairs for two single detectors across a day; (b) average flow vs. average density (per 5 min) from all the detectors across two different days; (c) a map of Yokohama and the study site; (d) average speed vs. average occupancy (Each of the 8 different stamps represents a different time period for two days – onset and decay of AM and PM peak for a weekend and a weekday)

If traffic behaves as a queue through a bottleneck, traffic flow departing the queue may not stay at its maximum if (a) vehicles in the queue could not travel fast enough so that the front of the following car could not reach the point of the front of the leading car in the time allotted the service rate (This is very slow traffic, and may properly be called hypercongestion, but in general traffic departing the front of the queue will be faster, as the shockwave that reduces speed (the wave of red brake lights) has moved to the back of the queue); (b) if the departure flow is affected by external sources; or (c) the bottleneck is not being fully served.

Examining traffic upstream of the bottleneck is interesting, but does not get to the root of the problem – the bottleneck itself. This view of hypercongestion is thus not inconsistent with the conclusion drawn by Small and Chu (2003) that the hypercongested region is unsuitable for use as a supply curve in congestion pric-

ing analyses. We further suggest that the use of "supply curves" is often inappropriate, given the non-stationary nature of congestion.

We now show that the traditional average cost vs. demand curve (introduced by Pigou and applied in most of the marginal-cost models) does not provide an accurate representation of congestion when traffic conditions are not stationary. Consider a region of a city, e.g. the part of Downtown Yokohama shown in Fig. 1c, which traffic state is described by properties (i) and (ii) of the first paragraph of this section. Then the state of the system is governed by the mass conservation equation (Daganzo 2007)

$$\frac{dn}{dt} = I'(t) - o(n(t)), \tag{1}$$

where n is the accumulation (number of the vehicles in the system), $I'(t)$ is the input rate (inflow) to the system at time t, and o is the total outflow from the system as a function of accumulation. This equation simply explains that traffic systems are dynamic and to estimate the state of the system at time t (and then the average travel time through curve in Fig. 1d), the knowledge of the input flow is not sufficient, but boundary conditions are needed, i.e. the state of the system at a prior time t'. Thus, a traffic model that estimates the average travel time based on a specific demand-cost curve ignores not only variations in the demand, but more importantly that this travel time will be different if the initial state of the system is uncongested, near maximum flow conditions or in the congested regime.

To further explore this issue we analyze here some additional data from the Yokohama experiment and we present the results. For a whole day we calculated per 5 min the total input flow entering downtown (vehicles crossing the dashed line in Fig. 1c), the average network flow inside the region (vehicle miles traveled per unit time over total network length) and the average speed and we plotted the results for different times of a day, shown in Fig. 2. Pairs of total flow vs. pace (1/speed) rely in a well-defined curve (Fig. 2b), which is not affected by the different origin-destination pairs and demand variations across a day, i.e. it can describe a dynamic traffic system. Nevertheless, the input flow vs. pace curve not only has significantly more scatter, but also successive points follow different paths during the onset and offset of congestion in the morning and evening peak. Thus, in this paper we will proceed using the tool of the MFD to describe congestion dynamically and derive efficient pricing policies.

It is clear that the result of an efficient control policy (congestion pricing in our case) should not allow the system to reach states on the decreasing branch of the MFD (Fig. 1b). In which state of the MFD a city should be operate is a policy decision. For example, state R in Fig. 1b is a more reliable and less equitable state, because the average speed is higher, but the system operates at a space-mean flow below the maximum. Thus, fewer people (presumably those with a higher value of time) pay the pricing charge and travel in the rush hour. State E is more equitable (higher flow) and has a slower speed, but the total welfare may be smaller depending on the distribution of the value of time and costs of schedule delay within the

population. The model developed below provides the necessary tools to analyze the trade-off between reliability and equity.

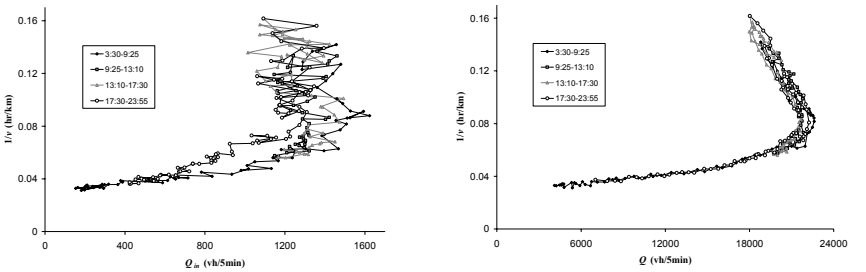

Fig. 2. Loop detector data from Yokohama: a) input flow vs. average pace (1/speed) pairs across a day; b) Average network flow vs. average pace pairs for the same day (Data resolution is 15min)

3. The Recurring Congestion Problem for a Network (No-toll Case) - Model Formulation and Equilibrium Solution

The purpose of this study is to investigate the recurring congestion problem (such as for the morning commute) at a network level for an urban area. The main difference between our problem and the one for a single bottleneck is that capacity (maximum flow in our case) is a function of average density during congested conditions. For a single bottleneck, queues form when demand is higher than capacity and the system operates at capacity until queues clear. At the network level, as more vehicles enter the system (accumulation on the right hand side of point E in Fig. 1b) the average network flow and the total outflow of the system decrease and, as a result, the effect of congestion is more significant. The consequence of this property for urban networks is that the estimation of equilibrium and the optimum toll are not straightforward and the analytical model should be refined. As we will show, many regularities of the single bottleneck model also apply in networks.

The formulation of the problem is as follows: It is assumed that a group of travelers wish travel through a network and reach their destination (e.g. work) on time. If $Z(t)$ denotes the cumulative number of travelers that wish to reach their destination by time t, then the demand rate, $Z'(t) \equiv dZ(t)/dt$, exceeds the maximum network outflow, γ, only during some time interval $[t_1, t_2]$ and $Z(t)$ is an S-shaped function (first derivative is always positive and second derivative is zero only once). Nevertheless, demand decreases below γ after t_2 and eventually everybody reaches his destination. Commuters are assumed to be aware of traffic

conditions day-after-day and they choose their departure time to minimize their individual total trip cost. This cost has two components, (i) the delay, *w*, because conditions are not free-flow and demand is high and (ii) a schedule penalty because commuters arrive at their destination early or late. The schedule penalty function, $c_s(\cdot)$ as in most of such studies is assumed to be piecewise linear, where t^* is the desired arrival time, and *e* and *l* are positive earliness and lateness rates,

$$c_s\left(t^* - t\right) = \begin{cases} -l\left(t^* - t\right), & if\ t^* \le t, \\ e\left(t^* - t\right), & if\ t^* > t. \end{cases} \tag{2}$$

The queue discipline is first-in-first-out (FIFO). The problem is, given $Z(t)$, to determine a curve $I(t)$, the cumulative number of departures from home at time *t* (inflow into the congested region), and a corresponding stable choice of arrivals at the final destination, $A(t)$, such that no traveler has an incentive to deviate from his choice and find a departure time giving less total trip cost than the one experiences. Under a FIFO queueing discipline the travel delay for a given vehicle arriving at its destination at time *t*, $w(t)$, is the horizontal distance between curves *I* and *A*, while his scheduled delay, $s(t)$ is the horizontal distance between curves *A* and *Z*. We assume that the average travel time, τ, and the total outflow, *o*, are a function of the system accumulation *n*, i.e. the network is described by an MFD and average trip length, *L*, is constant over time (Average network speed is by definition vehicle-miles traveled per unit time over average accumulation, while vehicle-miles is estimated during steady-state conditions as the product of outflow and trip length):

$$\tau(n) = \frac{n}{o(n)L}. \tag{3}$$

Travel delay is defined as the additional travel time to the travel time for a maximum flow, γ,

$$w(n) = \tau(n) - \tau(\min(n, n_0)), \tag{4}$$

where n_0 is the accumulation of the system when maximum outflow, γ, occurs. The assumption here is that commuters when conditions are not congested $(n < n_0)$ will arrive at their destination on time, which means that they have no incentive to experience schedule delay to improve their total trip cost. By observing real data from Yokohama, it turns out that this is a logical assumption for real cities because speed changes with time outside the rush hour are smooth enough to make the decision of travelling at higher speed and experience an earliness penalty, not worthwhile (the small time saved traveling by arriving early is outweighed by the early arrival penalty). The consequence of this assumption is that curves *A*, *I* and *Z* coincide when $n < n_0$; thus, the problem can be expressed in an elegant form as shown in Fig. 3. (There is an analogy with the assumption of zero travel time until travelers reach the bottleneck for the Vickrey's model. It is also useful

when commuters, who leave in the suburbs travel a free-flow distance before reaching the congested network, and this time is omitted from our model.)

For a given function $Z(t)$ of the shape of Fig. 3 there should be some (unknown for now) time t_0 when system reaches the state of maximum flow for the first time. Any traveler prior to time t_0 can reach his destination with zero travel and schedule delay. Again if the congestion vanishes after t_3 the traveler will also choose to have zero cost. This means that $A(t) = Z(t) = I(t)$ for $t \le t_0$ and $t \ge t_3$. In contrast with the single bottleneck model, given t_0, the estimation of t_1 is not trivial as $dA/dt \le \gamma$ for $t_0 < t < t_3$. The problem is to determine t_0, t_3 and $n(t)$ or $A(t)$, as by knowing the one we estimate the other through equation (1).

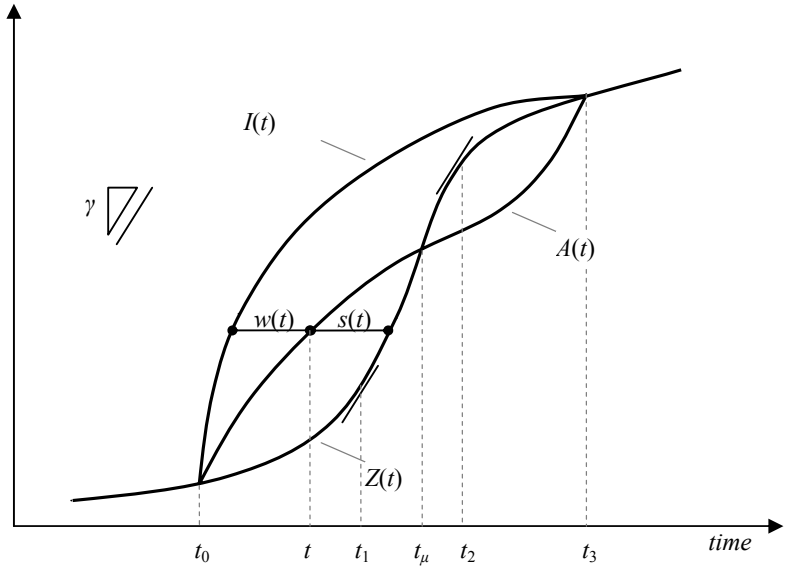

Fig.3. Cumulative number of travelers who enter the network, $I(t)$, arrive at their destination, $A(t)$, or have work starting times, $Z(t)$, less than t. $(A'(t_0) = A'(t_3) = \gamma = Z'(t_1) = Z'(t_2))$

If c_w is the shadow cost of travel delay, the trip cost for a traveler that reaches his destination at time t, but has a desired time t^* is

$$c(t,t^*) = w(n(t))c_w + c_s(t^* - t)\,. \tag{5}$$

Given that (i) $c(t_0,t_0) = c(t_3,t_3) = 0$ and (ii) $c(t,t^*) > 0$ for all $t,t^* \in (t_0,t_3)$, a necessary condition for a minimum $c(t,t^*)$ for a specific traveler is

$$\frac{\partial c(t,t^*)}{\partial t} = w'(n(t)) \cdot n'(t)c_w - c_s = 0\,. \tag{6}$$

The traveler that experiences the maximum travel delay arrives on time, at time t_μ. Otherwise, he could start his trip slightly earlier (later) if he was late (early) and decrease both his schedule and travel delay, i.e. his total trip cost. (The single bottleneck model reaches the same conclusion and queueing delay also has a triangular shape.) The ordinary differential equation in equation (6) does not contain the desired arrival time t^* and it can be solved simultaneously for all travelers, as they all see the same travel delay function $w(n(t))$. Thus, the travel delay will have a triangular shape with time, increasing from t_0 to t_μ with a slope e/c_w and then decreasing from t_μ to t_3 with slope $-l/c_w$, as shown in Fig. 4:

$$w(n(t)) = \begin{cases} \dfrac{e}{c_w}(t - t_0), & if\ t \le t_\mu, \\ -\dfrac{l}{c_w}(t - t_3), & if\ t > t_\mu. \end{cases} \tag{7}$$

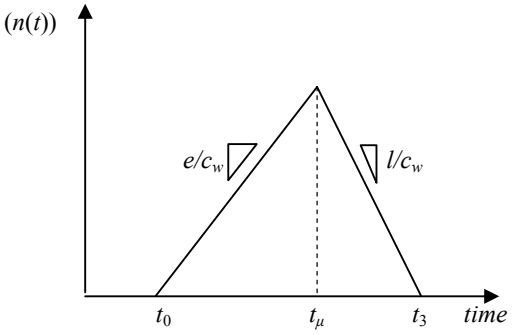

Fig.4. Delay for travelers arriving at their destination at time *t*.

Now, it is straightforward to estimate $n(t)$ and then $A(t)$ through equation (8). Empirical observations show that $w(n)$ is an one-to-one function (monotonically decreasing for $n > n_0$), so it is invertible.

$$n(t) = \begin{cases} w^{-1}\left(\dfrac{e}{c_w}(t - t_0)\right), & if\ t \le t_\mu, \\ w^{-1}\left(-\dfrac{l}{c_w}(t - t_3)\right), & if\ t > t_\mu. \end{cases} \tag{8}$$

Also, it is easy to show that $A(t)$ and $Z(t)$ intersect exactly once in the interval (t_0, t_3), at time t_μ. The slope of $A(t)$ depends on the value of $n(t)$, which has a similar shape with $w(n(t))$, increasing in $[t_0, t_\mu)$ and then decreasing in

$(t_\mu, t_3]$. Thus, $A(t)$ is also an S-shaped function (where the first derivative is always positive and second derivative is zero only once) and cannot intersect with $Z(t)$ more than once given that $A(t) = Z(t)$ for $t = t_0, t_3$.

To have a complete solution we need to determine t_0, t_μ and t_3. We will present here a constructive solution as an analytical solution for a general $Z(t)$ function looks very tedious. The implementation consists of the following steps:

1. Choose a $t_0 < t_1$. Estimate $n(t)$ using equation (8) and then $o(n(t))$ through equations (3) and (4).
2. Construct $A(t) = \int_{\tau_0}^{\tau} o(n(\tau) d\tau$ and find t_μ, t_3 as the intersection points of $A(t)$ and $Z(t)$. Estimate $\rho = (t_3 - t_\mu)/(t_\mu - t_0) - e/l$.
3. If $\rho = 0$, this is the right solution. Otherwise repeat steps 1 and 2 by choosing a smaller t_0 if $\rho > 0$ or a larger t_0 if $\rho < 0$.

We should note that Small and Chu (2003) solved a similar formulation of the no-toll network equilibrium under many simplified assumptions: (i) all travelers have the same desired arrival time t^*, (ii) demand is constant in time interval $[t_1, t_2]$ and (iii) speed is a linear function of network accumulation. They did not obtain the regularities obtained in our paper.

4. Cordon-based Optimal Pricing for a Network

4.1 An Optimal Fine Toll to Reduce Congestion

We provide here an analytical derivation of the social optimum using a time-dependent ("fine") toll. As travel delay is a deadweight loss, we are looking for this toll that will minimize total travel time, i.e. travel delay, as defined in equation (4), will be zero for all travelers. Thus, $A(t) = I(t) \ \forall t$ and we are interested only in curves *A* and *Z*. Vickrey (1969) pointed out that if we charge a time-dependent toll equal to the time spent in queue (travel delay in our case) travelers would arrive at the bottleneck (enter the network) at such times that queues (travel delay) would not form. The effect of this is that "one could convert the worthless expense of queueing into money" (Newell 1987). Also, Arnott et al. (1990) described that the length of the rush hour will be the same in the toll case and the social optimum. But, this does not hold for the network model.

The important difference between a single bottleneck with constant capacity and a network is that $t_0 \neq t_0^{toll}$, $t_3 \neq t_3^{toll}$, $t_\mu \neq t_\mu^{toll}$ because of time-varying throughput (During congestion an active bottleneck operates at capacity if there is no restriction downstream, but a network is described by an MFD, which means that the rate users reach their destination decreases for high congestion levels). The advantage of applying the toll is not only that delays disappear but also the length of the rush hour becomes shorter. Also, the average optimal fine toll is smaller than the average delay cost (Arnott et al. (1993) shows that these two quantities are the same for the single bottleneck). This means that the savings in travelers' delay are significantly higher for the network model. Mathematically speaking, we show here how the values for the optimal toll $T(t)$ and the beginning and ending times t_0^{toll} , t_3^{toll} are derived.

First, we are looking for the smallest possible time-dependent toll that will keep travel delays, as per equation (4), at zero. Obviously, after the enforcement of this toll, system should operate at maximum outflow, γ. Otherwise, with a slightly smaller toll, we could allow more people to enter the system with zero delay penalty, as per equation (4). (In the next section we analyze the potential travel time savings by applying a stricter toll.) Fig. 5 describes our problem with curves for cumulative actual arrivals, $A_{toll}(t)$, and desired arrivals, $Z(t)$.

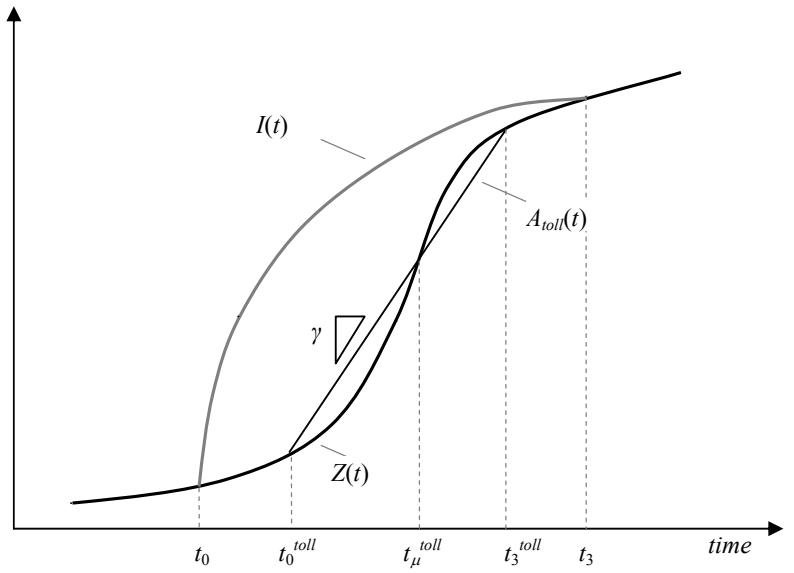

Fig. 5. Cumulative number of travelers who arrive at their destination, $A(t)$, or have work starting times, $Z(t)$, less than t for the optimum fine toll (zero delay).

Second, $t_0^{toll} < t_1$ and $t_2 < t_3^{toll}$, i.e. to eliminate travel delay the time window for which desired arrival rates are higher than outflow γ should be a subset of the time window for which the toll is applied. The total cost for a traveler with desired arrival t^*, who arrives at destination at time t is

$$c_{toll}(t,t^*) = T(t) + c_s(t^* - t). \tag{9}$$

Travelers with desired arrivals at t_0^{toll} (or t_3^{toll}) experience zero earliness (or lateness) penalty. Otherwise, they would have the incentive to begin their trip slightly earlier (or later) and decrease their total trip cost. For the same reason $T(t_0^{toll}) = T(t_3^{toll}) = 0$. If $T(t)$ is a continuous and differentiable function in the interval $[t_0^{toll}, t_3^{toll}]$, then a necessary condition for a specific traveler to minimize his total trip cost is $\partial c_{toll}(t,t^*)/\partial t = 0$. Solving for $T(t)$ we get:

$$T(t) = \begin{cases} \dfrac{e}{c_w}\left(t - t_0^{toll}\right), & if\ t_0^{toll} \le t \le t_\mu^{toll}, \\ -\dfrac{l}{c_w}\left(t - t_3^{toll}\right), & if\ t_\mu^{toll} \le t \le t_3^{toll}, \\ 0, & \text{all other times.} \end{cases} \tag{10}$$

Estimation of t_0^{toll}, t_μ^{toll} and t_3^{toll} is less tedious now. We can apply the same procedure with the no-toll case described in the end of Section 3.1. The only difference is that $A(t) = Z(t_0^{toll}) + \gamma t$.

While savings for travelers in the single bottleneck level was 0 (toll=waiting time), savings for travelers who pay toll are positive and consist of (i) savings in travel delay, which are higher than the toll paid and (ii) savings in the schedule delay. If $\delta(\gamma)$ is the duration of the toll period for outflow γ, and $\delta(\text{no toll})$ the length of the rush hour in the no-toll case, total savings for all travelers (including the cost of tolls) are

$$\Delta\$ = \frac{1}{2}\left(\delta(\text{no toll})^2 - \delta(\gamma)^2\right)\frac{\gamma l e}{l+e}. \tag{11}$$

Total savings are equal to the total travel delay in the no-toll case (Area of triangle in Fig. 4) minus the total cost of toll during period $\delta(\gamma)$. Equation (11) omits savings in schedule delay because the estimation is tedious. In any case, $\Delta\$$ are direct savings that can be estimated even if $Z(t)$ is unknown (which is consistent with reality wherein transportation agencies can measure travel delay but not schedule delay). Section 4.3 will present an example using some real data from Yokohama.

4.2 Policy Implications: Equity vs. Reliability

In the previous sections we omitted the fact that traffic states with outflow less than γ, will experience higher average speeds and therefore lower travel times. Although this assumption does not change the equilibrium in the no-toll case, it ignores potential savings in travel delay if a higher toll than the one estimated in Section 4.1 is applied. A higher toll will push the system to operate at an outflow $\gamma' < \gamma$ during the toll period. This also implies that we should use a smaller value for n_0 in equation (4). For example, state R in Fig. 1b is a more reliable and less equitable state, because the average speed is higher (and more likely to be stable), but the system operates at a space-mean flow below the maximum. Thus, fewer people (presumably those with a higher value of time) pay the toll and travel in the rush hour. State E is more equitable (with a higher flow) but has a slower speed, so the total welfare may be smaller depending on the distribution of the value of time within the population. Also, state R is more reliable than state E, because as $Z(t)$ and schedule penalty rates are unobservable quantities, a toll may not be efficient at all times, and certain tolls allow the system to reach congested states, where $n > n_0$. In which state of the MFD a city should operate is a policy decision. We estimate now what is the change in the total cost because of a stricter toll.

Fig. 6 plots the cumulative arrivals for different capacities with $\gamma' < \gamma$. But, travel time for outflow γ' and γ are different and this has been omitted from the graph, as for cumulative actual arrivals, $A_{toll}^{\gamma'}(t)$, we have chosen a smaller value for n_0 in equation (4). In the extreme case we choose outflow $\gamma'' = (Z(t_3) - Z(t_0))/(t_3 - t_0)$, this implies that (i) the toll period is equal to the rush hour period for the no toll case and (ii) time dependent toll equals to the cost equivalent of the travel delay $(\Delta\$ = 0)$, i.e. the same conclusions with the single bottleneck model. In the latter case, the benefit for the traveler is almost zero (only a small reduction in schedule delay) while the benefit for the city (toll operator) is higher. That is, tolls can be set higher to be profit maximizing or lower to be welfare maximizing. Traditionally welfare calculations have been dominated by travel time and tolls; however, welfare maximization is itself ambiguous, and tolls may be higher or lower depending on how welfare is calculated, e.g. does it account for schedule delay, reliability, inter-personal equity, etc. Our case does not allow for demand elasticity, which poses a natural future extension, so a profit maximizing toll here would be infinite.

If γ' is close to γ function $Z(t)$ is smooth in the beginning and the end of tolling period, we can assume that t_μ^{toll} is approximately the same in the two cases (this assumption makes the calculations less tedious; the exact solution is not presented here). The change in the total cost $\Delta C(\gamma, \gamma')$ is the sum of positive changes in to-

tal schedule delay $\Delta S(\gamma,\gamma')$ and negative changes in total travel time $\Delta T(\gamma,\gamma')$. Schedule delay changes occur because the length of tolling period increases and as the toll is higher, the length of the earliness and lateness period increase as well. Travel time changes occur because people are travelling faster during the tolling period $[\delta(\gamma') \cong \delta(\gamma)\gamma/\gamma']$.After some manipulations and first order approximations we have that:

$$\Delta S(\gamma,\gamma') \cong \frac{1}{2}\left(t_3^{toll} - t_0^{toll}\right)^2 \gamma^2 \left(\frac{1}{\gamma'} - \frac{1}{\gamma}\right)\frac{le}{l+e} > 0\,, \tag{12a}$$

$$\Delta T(\gamma,\gamma') \cong \left(t_3^{toll} - t_0^{toll}\right)^2 \gamma\left(\tau(o^{-1}(\gamma')) - \tau(o^{-1}(\gamma))\right)c_w < 0\,. \tag{12b}$$

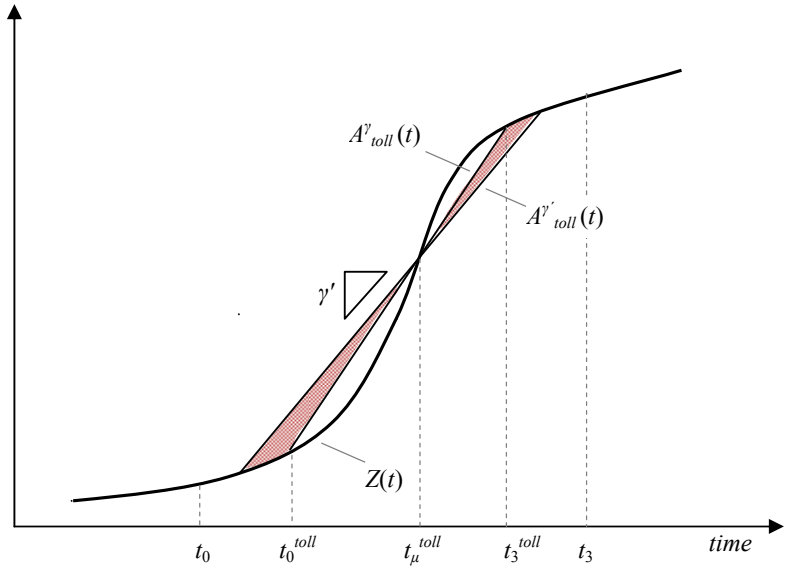

Fig.6. Cumulative number of travelers who arrive at their destination, or have desired arrival less than *t* for different values of capacities $(\gamma' < \gamma)$ with a fine toll (zero delay).

Thus, it depends on the values of the related parameters if such a decision is beneficial for the average traveler. By combining equations (12a) and (12b) a higher toll maybe beneficial for the population if ratio ω is smaller than 1, e.g. when function $\tau(n)$ is sharp on the left hand side of n_0, or schedule penalty rates are low.

$$\omega = \frac{\frac{1}{2}\left(t_3^{toll} - t_0^{toll}\right)^2 \left(\frac{\gamma}{\gamma'} - 1\right)\frac{le}{c_w(l+e)}}{\tau(o^{-1}(\gamma)) - \tau(o^{-1}(\gamma'))}\,. \tag{13}$$

5. An Example based on Field Observations

We apply now the models developed in Sections 3 and 4 to a test site (Downtown Yokohama, Japan). The part of downtown Yokohama examined in this paper is approximately a 10 km^2 triangle, which includes streets of various types, with closely spaced signalized intersections (~100-300m), and a few elevated freeways. Yokohama's center is congested during peak hours with average speeds less than 10 kilometers per hour observed for extended periods during morning and evening peak in weekdays in the arterial network. A previous study has analyzed a combination of fixed (detectors) and mobile (GPS in vehicles) sensors data and estimated relations between average traffic variables during different days in resolution of 5 min (like the ones presented in Figs. 1 and 2, i.e. accumulation, average travel time, outflows, input flows etc). It also provided a method to estimate the accumulation of the network with time. This study verified that (i) an MFD (relating average flow and density) can describe aggregate traffic in this site and (ii) the ratio of network flow over outflow (which is linearly related to the average trip length-see equation (3)) is an invariant during a day and across days, i.e. the traffic model used to describe travel behavior and derive equilibrium conditions for the recurring congestion problem of Sections 3 and 4 provides a realistic representation of macroscopic traffic conditions in the area. For a detailed description of the experiment the reader can refer to Geroliminis and Daganzo (2008).

Fig. 7 shows the estimation of accumulation during the morning peak for the study site using real data for one weekday. The point of the MFD with maximum outflow is the pair $(n_0, \gamma) = (7600vh, 785vh/5\,\text{min})$. We also fitted a curve in travel time vs. accumulation data for $n > n_0$ and a linear function provided an adequate fitting. Trip duration (in minutes) is $\tau(n) = 2.52 \times 10^{-3} n - 8.08$ $(R^2 = 0.79)$.

An interesting observation is that we can test some of the theoretical results of Section 3, without requiring knowledge of the $Z(t)$ and values for *l* and *e*. Actually, we show that we can estimate these by looking at this traffic data. Fig. 7 shows that we can approximate quite well the onset and offset of congestion as a piecewise linear function of accumulation with time. (We do not consider accumulation in the tails of the peak period as traveler behavior may be different). As $\tau(n)$ is linear and $n(t)$ has a triangular form, $w(t)$ will have the same (triangular shape) with $n(t)$. Furthermore, the ratio of slopes for the onset and the offset of congestion will be the same for travel delays and accumulation. Thus, we observe with real data that (i) travel delay, as per equations (4) and (7), is a realistic approximation of our study site and (ii) we can estimate the ratio of earliness over lateness rates $e/l \cong 3139/6123 = 0.51$, which value is in accordance with an earlier study (Small 1982).

We now derive a cordon-based congestion pricing for downtown Yokohama based on the theoretical results of Section 4. Note from Fig. 7 that we can extract

the values for t_0, t_3 and t_μ; the beginning, the end of the peak hour and the time when delay is maximum. Thus, we can reproduce $A(t)$ and $I(t)$ curves described in Fig. 3. This is shown in Fig. 8 and the area between *A* and *I* curves is the total travel delay, as per equation (4), during the morning peak period.

Given the value of γ, we could construct an optimal fine toll solution, if we knew $Z(t)$. Neverheless, $Z(t)$ is not easily observable. To simplify matters we assume that all the desired arrivals are uniformly distributed in an interval *D*, subset of the interval $[t_0, t_3]$, i.e. $Z(t)$ is flat outside *D* and has a constant slope, *s*, in *D*. Under this condition, one can easily show that $t_\mu^{toll} = t_\mu$. Thus, if we know the shadow price of travel delay c_w, we can evaluate the optimal fine toll, as per equation (10), because *e*/*l* is already known from Fig. 7.

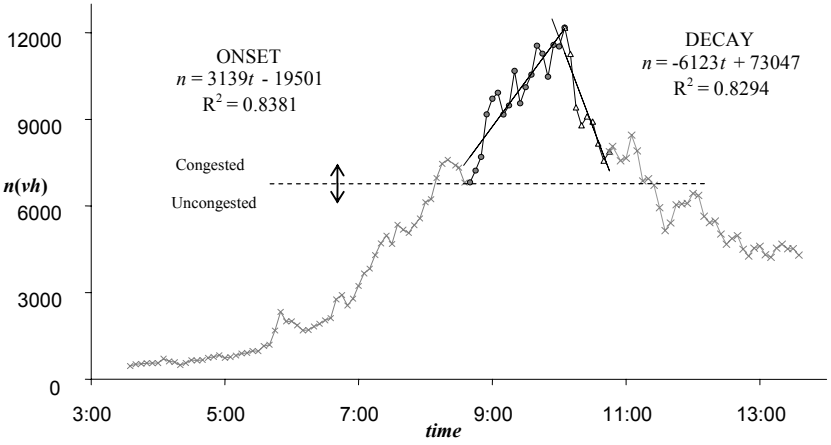

Fig. 7. Number of vehicles in downtown Yokohama during the day of Dec. 14, 2001.

We now treat equity and reliability issues, by evaluating the optimal fine toll for different values of maximum flow, i.e. in what mobility level the study city should choose to operate. For this analysis we assume that $s = 1000 vh/hr$, 25% higher than γ. Besides γ, we estimate the optimal fine toll for $\gamma'' = (Z(t_3) - Z(t_0))/(t_3 - t_0)$ and a value $\gamma < \gamma' < \gamma''$. Fig. 8 also shows the actual arrivals in destination $A_{toll}^{\gamma}(t)$, $A_{toll}^{\gamma'}(t)$ after the implementation of tolls according to the model. Table 1 summarizes the results for the no-toll case and the optimal fine toll for γ, γ' and γ'' in terms of schedule delays, travel delays and tolls paid and duration of the toll period (length of peak period in the no-toll case).

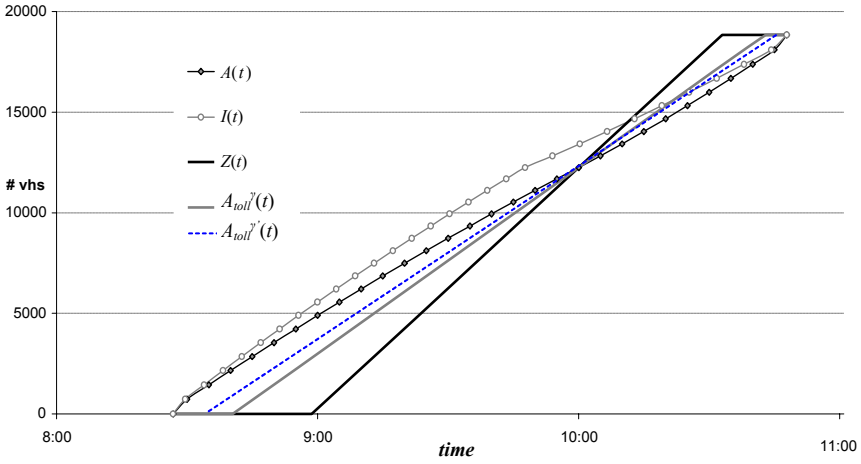

Fig. 8. Application of cordon based pricing in the downtown Yokohama for maximum outflow γ (equity) and a smaller value γ' (reliability). Equilibrium solutions with and without tolls $(s = 1000vh/hr, c_w = \$20/hr)$.

Table 1. Comparison of schedule, travel delays and tolls for different cordon-based fine tolls applied in Yokohama.

	No-toll	**Toll for γ**	**Toll for γ'**	**Toll for γ''**
Toll duration (hr)	2:21	2:00	2:09	2:21
Start time	8:27	8:41	8:35	8:27
End time	10:48	10:41	10:44	10:48
Total Travel Delay(vh-hr)	1792	0	-283	-581
Av. Travel Delay(min/vh)	5.94	0	-0.9	-1.85
Total earliness (vh-hr)	3502	1713	2309	3105
Total lateness(vh-hr)	985	496	669	900
Aver. earliness (min/vh)	17.2	8.4	11.3	15.2
Aver. Lateness (vh/min)	9.0	4.5	6.1	8.2
Aver. Toll ($)	0	1.69	1.81	1.98
Aver. Toll (travel min)	0	5.06	5.43	5.94
Max. outflow (vh/5min)		784.9	730.1	668.0

In the case of cordon pricing for the maximum possible outflow, the travel delays are eliminated, while schedule delays are about half of the no-toll case (8.4min and 4.5min average earliness and lateness vs. 17.2min and 9min respectively). The average travel delay savings are about 6min/vh which is 75% of the average travel time at outflow γ. (Note that these travel times are relatively small because the study area is about 10 km^2 and we consider only the travel inside the

region, i.e. the average trip length L, as per equation (3), is the average distance traveled inside the region per trip completion.)

By charging a slightly higher toll ($1.81 instead of $1.69) each traveler saves about 1min of travel delay but schedule delays increase more (2.9min the earliness and 1.6min the lateness). Thus, it depends on the relative ratio of e/c_w which policy is better. For the extreme case of γ'', schedule delays increase at a higher rate.

The parameters that affect the described pricing scheme are the *e*/*l* ratio, the shape of the MFD, the value of time c_w and the $Z(t)$ function. As c_w and $Z(t)$ are not easily observable quantities, a city can identify the optimum toll by a trial-and-error procedure where the toll is modified based on the state of the city in the MFD (higher toll when $n > n_0$).

6. Discussion and Future Work

In this paper, we argued and verified with results from a real experiment that the standard economic models of marginal cost cannot describe traffic congestion in networks during time-dependent conditions. Then, we extended Vickrey's model of peak-hour for a single bottleneck, in the case of a network and we estimated a cordon-based fine toll to reduce congestion. The main difference between these two models is that for a single bottleneck, queues form when demand is higher than capacity and the system operates at capacity until queues clear. On the contrary, at the network level, maximum outflow is a function of average density during congested conditions and as more vehicles enter a congested system the average network flow and the total outflow decrease. We also examined some policy implications of the cordon-based pricing to treat equity and reliability issues, i.e. in what mobility level a city should choose to operate. We also applied the model to develop a pricing scheme for downtown Yokohama, without requiring accurate prediction of unobservable quantities.

One interesting result was that by applying an optimal toll not only delays disappear, but also the length of the rush hour shortens. Also, the savings in travel and schedule delay are much higher than the total amount of toll paid and this is true for every user. This scheme is Pareto-efficient as everybody is better-off by this policy and nobody can improve his situation by entering the system at a different time. Another interesting result is that total schedule delay costs are of the same order of magnitude as travel delay costs. Thus, models that ignore this significant cost for a traveler may yield to erroneous conclusions. While this model provides a simplified description about travelers' choices, the effect of congestion in their decisions and cordon-based pricing schemes to improve cities mobility, some extensions are needed.

First, the model developed in this paper is macroscopic for both travelers (demand) and the network (supply), each was based on statistical distributions. One could apply the same ideas in a microscopic behavioral micro-simulation (agent-

based) framework with individual travelers collectively creating congestion which takes the form of the MFD. Individual travelers will vary in the desired arrival time, early and lateness penalty, willingness to tolerate variability or unreliability in travel times, demand elasticity (willingness to not travel by car), and tolerance for delay. Use of the MFD may enable sketch-planning of cordon or area-based congestion pricing programs without specific network routing and network simulation results. Coupling the MFD with agent-based models would allow the examination of equity and efficiency effects of alternative pricing policies.

More empirical examples to justify the appropriateness of the MFD are in progress. We try to understand what types of networks and under what conditions experience an MFD with small scatter. Also, a fine toll is not always easy to implement. We are interested in finding the equilibrium in the case of coarse tolls for the network model. Arnott et al. (1993) solved this problem for a single bottleneck. The solution is not straightforward because $A(t)$ will not be S-shaped, as travel delays are not eliminated.

As well, it is desirable to apply these concepts in the case of elastic demand, where people not only change the starting time of their trip, but also can decide not to travel if e.g. the toll is too high or the system is extremely congested. Also, a successful implementation of the proposed pricing scheme will shorten the length of the congested peak period. This may induce more people to drive by car and decrease the modal shift. A model that can describe the dynamic nature of this issue or policy decisions that will address the induced demand can have some research priority.

References

Anderson, D. and Mohring, H. (1997). Congestion costs and congestion pricing. *The Full Costs and Benefits of Transportation: Contributions to Theory, Method and Measurement*. Springer.

Arnott, R., de Palma, A. and Lindsey, R. (1990a). Departure time and route choice for the morning commute. *Transportation Research Part B,* 24, 209-228.

Arnott, R., de Palma, A. and Lindsey, R. (1990b). Economics of a bottleneck. *Journal of Urban Economics,* 27, 111-130.

Arnott, R., de Palma, A. and Lindsey, R. (1993). A structural model of peak-period congestion: a traffic bottleneck with elastic demand. *American Economic Review,* 83**,** 161-179.

Cassidy, M.J. and Bertini, R.L. (1999). Some traffic features at freeway bottlenecks. *Transportation Research Part B*, 33(1), 25-42.

Crozet, Y. and Marlot, G. (2001). Congestion and road pricing: where is the 'bug'? *The 9th World Conference on Transport Research*, Korea.

Daganzo, C.F. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck, *Transportation Science* 19 (1), 29-37.

Daganzo, C.F. (2007). Urban gridlock: macroscopic modeling and mitigation Approaches. *Transportation Research Part B,* 41, 49-62.

Geroliminis, N. (2007). *Increasing Mobility in Cities by Controlling Overcrowding*, Ph.D. thesis, University of California, Berkeley.

Geroliminis, N. and Daganzo, C.F., (2007). Macroscopic modeling of traffic in cities. *The 86th Transportation Reserarch Board Annual Meeting*, Paper no. 07-0413, Washington DC.

Geroliminis, N. and Daganzo, C.F. (2008). Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transportation Research Part B, 42*(9), 759-770.

Hansen, M. and Huang, Y. (1997). Road supply and traffic in urban areas: a panel study, *Transportation Research Part A*, 31(3), 205-218.

Hau, T.D. (1998). Congestion pricing and road investment. *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*. Cheltenham, Edward Elgar.

Lago, A. and Daganzo, C.F. (2007). Spillovers, merging traffic and the morning commute. *Transportation Research Part B*, 41(6), 670-683.

Levinson, D. (2002). *Financing Transportation Networks.* Edward Elgar Publishers, Northampton.

Levinson, D. and Odlyzko, A. (2008). Too expensive to meter: the influence of transaction costs in transportation and communication. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences,* 366(1872), 2033-2046.

Liu, L.N. and Boyce, D. (2002). Variational inequality formulation of the system-optimal travel choice problem and efficient congestion tolls for a general transportation network with multiple time periods. *Regional Science and Urban Economics*, 32(5), 627-650.

Marchand, M. (1968). A Note on Optimal Tolls in an Imperfect Environment. *Econometrica,* 36, 575-581.

Maruyama T. and Sumalee A. (2007). Efficiency and equity comparison of cordon- and cordon-based road pricing schemes using a trip-chain equilibrium model. *Transportation Research Part A,* 41, 655-671.

Newell, G. (1987). The morning commute for non-identical travelers. *Transportation Science*, 21(2), 74-88.

Pigou, A.C. (1920). *Wealth and Welfare*. 1st edition.

Small, K. (1982). The scheduling of consumer activities: work trips. *The American Economic Review,* 72(3), 467-479.

Small, K. and Chu, X. (2003). Hypercongestion. *Journal of Transport Economics and Policy,* 37(3), 319-352.

Smith, M.J. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science,* 18(4), 385-394.

Vickrey, W. (1963). Pricing and resource allocation in transportation and public utilities. *The American Economic Review*, 53(2), 452-465.

Vickrey, W. (1969). Congestion theory and transport investment, *American Economic Review,* 59, 251-260.

Wardrop, J. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers,* 1(2), 325-362.

Yang H. and Huang H.J. (2005). *Mathematical and Economic Theory of Road Pricing*. Elsevier Science Inc, New York.

Yang, H. and Huang, H.J. (1998). Principle of marginal-cost pricing: how does it work in a general network. *Transportation Research Part A,* 32, 45-54.

Yang, H. and Meng, Q. (1998). Departure time, route choice and congestion toll in a queuing network with elastic demand. *Transportation Research Part B*, 32(4), 247-260.

Yang, H. and Lam, W.H.K. (1996). Optimal road tolls under conditions of queuing and congestion. *Transportation Research Part A,* 30, 319-332.

Zhang, L. and Levinson D. (2004). Some properties of flows at freeway bottlenecks. *Transportation Research Record,* 1883, 122-131.